Rejoinder: Remaining Challenges in Investigating Grade Retention Effectiveness

Machteld Vandecandelaere

Centre for Educational Effectiveness and Evaluation, University of Leuven

Stijn Vansteelandt

Department of Applied Mathematics, Computer Science and Statistics, Ghent University

Machteld Vandecandelaere

Centre for Educational Effectiveness and Evaluation,

University of Leuven, Leuven, Belgium

Machteld.Vandecandelaere@kuleuven.be

Stijn Vansteelandt

Department of Applied Mathematics, Computer Science and Statistics

Ghent University, Ghent, Belgium

Stijn.Vansteelandt@ugent.be

Correspondence concerning this article should be addressed to Machteld Vandecandelaere, The Education and Training Research Group, Centre for Educational Effectiveness and Evaluation, Dekenstraat 2 (pb 3773), 3000 Leuven, Belgium.

E-mail: Machteld.Vandecandelaere@kuleuven.be

To cite this article: Vandecandelaere, M., & Vansteelandt, S. (2016). Rejoinder: Remaining Challenges in Investigating Grade Retention Effectiveness. *Multivariate Behavioral Research.* Advance online publication. doi: 10.1080/00273171.2016.1229171

Abstract

This rejoinder, in response to the commentaries of Steiner, Park and Kim (this issue) and Reshetnyak, Cham and Hughes (this issue), discusses remaining challenges in grade retention research. First, a same-age comparison assumes that the instruments used in different grades measure ability equally well. We discuss the importance of evaluating the properties of the scaling process to address whether this assumption has been met. Second, we discuss issues in the selection of covariates to be included in the weights. Third, we discuss the unconfoundedness assumption and the problem of remaining imbalance. Finally, we provide an empirical illustration showing that studying grade retention effectiveness comes with multiple methodological decisions that are rooted in a bias-variance trade-off.

Keywords: marginal structural models, unconfoundedness assumption, grade retention, bias-variance trade-off

Rejoinder: Remaining Challenges in Investigating Grade Retention Effectiveness

We thank all commentators for valuable commentaries that provide useful insights into the challenges regarding grade retention research and methods for investigating time-varying treatments more generally. Building on these commentaries, we organize this rejoinder around remaining challenges involved at different stages of research design and analysis in investigating grade retention effectiveness.

Same-age Comparison: Comparing Test Scores in Different Grades

Steiner, Park and Kim (this issue) give a clear formalization of same-grade and same-age comparisons using the potential outcomes framework. Although same-grade comparisons are useful for comparing the performance of retained and promoted students at a specific grade, same-age comparisons are more suited to assess the double-dose effect of repeating a grade. Steiner et al. (this issue) correctly point out that a thorough discussion of whether the assumptions underlying the selected comparison strategy are met is essential in assessing the meaning and credibility of estimated effects. In our study, we used a same-age comparison. This implies, for example, that comparing retained and promoted children at the age of ten involves comparing children in fourth and fifth grade respectively. To test their mathematics achievement, different instruments are used, adapted to the specific grade (i.e., the test in fifth grade was more difficult than the test in fourth grade). The key assumption is that both tests measure student's math ability equally well and the same construct must be measured by the tests. In view of this assumption, the items of both tests need to be vertically equated, meaning that the items of the different tests need to be situated on a common ability scale. Using Item Response Theory, this is possible by including a subset of anchor items in both tests. The anchor items are used to scale one test to the other. For vertical equation to be successful, the anchor items should have a high level of discrimination, meaning that the items can differentiate well

between students with a low and a high ability level. Furthermore, the difficulty and the level of discrimination of the anchor items should be the same for retainees and promoted children. We recommend checking the item characteristics for both groups and evaluating measurement invariance between the groups. For equating to be successful, the range of ability of the retained and the promoted group need to overlap. In other words, for an anchor item to be useful, each group should contain a number of students for which the item is not too easy and not too difficult. After vertically equating, a student should obtain the same ability estimate, regardless of the set of items used (for more details, see Baker (2001)).

Covariate Selection using Marginal Structural Models

Key to the creation of the weights is the selection of variables to be included in the treatment assignment model. Research and guidelines on which covariates should be included in propensity score models have been scattered across disciplines. Propensity score methods, as originally defined by Rosenbaum and Rubin (1983), are inspired by the features of a randomized experiment, in which the outcome is unknown. Therefore, it has been argued that the propensity score should be based on all covariates related to the treatment, irrespective of their relation to the outcome. However, as Reshetnyak, Cham and Hughes (this issue) note, selecting covariates regardless of the relationship with the outcome should be avoided. Research has demonstrated that the specific set of covariates plays a part in the amount of bias and variance of the estimated treatment effect. Every step toward better balance usually comes with an increase in variance (Golinelli, Ridgeway, Rhoades, Tucker, & Wenzel, 2012). This increase in variance presents a special problem when variables are included that that are predictive of treatment assignment but unrelated to the outcome (i.e., so-called instrumental variables). Instrumental variables inflate any remaining biases that may be present (Myers et al., 2011; Pearl, 2011). Some authors therefore recommend including all variables that are

thought to be related to the outcome in the propensity score model, regardless of their association with treatment assignment (e.g., Brookhart et al., 2006; Myers et al., 2011).

The selection of confounders forms an area of vigorous research in statistics. In our opinion, such selection should ideally be done with the aim of optimizing the quality of the treatment effect estimate (e.g., minimizing its mean squared error). This criterion has been proposed for simpler settings that involve a single treatment at a given time (van der Laan and Gruber, 2010; Vansteelandt, Bekaert and Claeskens, 2012). We foresee that the development of covariate selection strategies will soon have progressed sufficiently far to meet the complexity of our analyses, which involve time-varying treatments. For now, we recommend comparison of the results of the use of different covariate selection strategies as a type of sensitivity analysis. The more the results are in line with each other, the more robust the results are to the covariate selection process.

Unconfoundedness Assumption and Balance

As noted by Reshetnyak et al. (this issue), the standardized mean difference (SMD) is only one criterion for evaluating the balance properties for the observed covariates. Variances, percentiles, boxplots, quartile-quartile plots of the covariates and higher-order terms could also be compared. Yet, even if these balance diagnostics signal no evidence of imbalance, the treatment effect estimate is unbiased only to the extent that the treatment model is correctly specified and includes all relevant confounders. Sensitivity analysis might give an indication of the extent to which the estimates are prone to a violation of the unconfoundedness assumption. A popular approach, initiated by Cornfield in epidemiology (Gastwirth, Krieger & Rosenbaum, 1998), is to evaluate how strong an unobserved binary confounder U would need to be associated with the treatment and outcome to change the conclusions of the study. This method has recently been generalized to avoid imposing assumptions on the unmeasured confounder

(e.g. that the unmeasured confounder is binary) (Ding & VanderWeele, 2016; see also VanderWeele & Arah, 2011). For marginal structural models in particular, a sensitivity analysis strategy was proposed in Brumback, Hernán, Haneuse and Robins (2004), but has the disadvantage of demanding the postulation of difficult-to-interpret sensitivity parameters.

Recall that adjusting for observed confounders also adjusts for unobserved confounders to the extent that these are correlated with the observed ones (Stuart, 2010). In other words, bias due to unmeasured confounding only results from covariates that are unrelated to the observed confounders. Since we included multiple observed confounders, including pre-treatment measures of the outcomes, we are confident that our analyses have eliminated a substantial degree of confounding bias. However, we cannot exclude the possibility that some degree of unmeasured confounding bias remains.

Reducing Reality to Disentangle the Truth?

To understand the effects of grade retention, we made use of a rich longitudinal dataset and we used statistical modeling to reduce the data into manageable and interpretable entities. The research design and the analysis come with multiple decisions, each of which may involve a bias-variance trade-off. Imposing more modeling assumptions enables us to draw more precise conclusions, but with an increased risk that these conclusions will become distant from reality to the degree that the underlying assumptions do not hold. On the other hand, the more variance is allowed, the more uncertain the results are.

For the sake of an illustrative comparison, we compare in Table 1 the results of the current study (Study 2) with the results of an earlier study in which we used the same dataset and addressed the same research question but which entailed more assumptions and ignored some of the complexities originating from the presence of time-varying confounders (see Vandecandelaere, Schmitt, De Fraine & Van Damme, 2015). Table 1 gives an overview of the

contrast estimates, standard errors, and effect sizes in Year 1 and Year 6 between three treatment conditions: continuous promotion, kindergarten retention and first-grade retention. It is clear that the approach in Study 1 yielded the highest precision. The standard errors are substantially smaller compared to those in Study 2. Decisions that required additional assumptions in Study 1 were, for example, matching on time-fixed covariates only, and, modelling mathematics development as a curvilinear process. On the other hand, weighting on time-varying covariates, and modelling time as a multivariate response relaxed some of the assumptions. These decisions caused a loss of precision in the estimated treatment effects in the current study (Study 2); however, the obtained standard errors likely give a more honest reflection of the uncertainty in the retention effect estimates. In particular, in Study 1, the advantage for kindergarten repeaters compared to first-grade repeaters in Year 6 was significant, whereas in Study 2 it was not. Because our studies differed with respect to many more aspects of study design, identification of the exact causes of the differences in power is complex. Future simulation studies might give more insight into the sources of these differences.

Table 1

Contrasts, standard errors and effect sizes in study 1 (Vandecandelaere et al., 2015) and study 2 (Vandecandelaere et al., 2016)

		K-retention – No retention					K- retention – G1-retention			
		Estimate		*SE*	*ES*		Estimate		*SE*	*ES*
Study 1	Year 1	-12.68	***	0.69	1.82	Year 2	-2.62	***	0.68	0.42
Study 2		-11.91	***	1.44	1.33		-3.47	**	1.22	0.37
Study 1	Year 6	-5.15	***	0.87	0.62	Year 6	3.47	**	1.12	0.39
Study 2		-4.95	*	2.47	0.54		3.96		2.58	0.43

In sum, studying the effectiveness of grade retention comes with multiple conceptual and methodological challenges. The article and the commentaries give an overview of the pros and cons in choosing a comparison strategy and illustrate the importance of carefully evaluating

the assumptions underlying the comparison strategy and the balancing process. We recommend to cautiously consider the bias-variance trade-off and the implications of each step in deciding on the research design and analysis strategy in investigating grade retention effectiveness.

References

Baker, F. B. (2001). *The basics of items response theory*. College Park, MD: University of Maryland, ERIC Clearinghouse on Assessment and Evaluation.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*(12), 1149-1156. doi:10.1093/aje/kwj149

Brumback, B. A., Hernán, M. A., Haneuse, S. J., & Robins, J. M. (2004). Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine, 23(*5), 749-767. doi: 10.1002/sim.1657

Ding, P. P., & VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology*, *27*(3), 368-377. doi: 10.1097/EDE.0000000000000457

Gastwirth, J. L., Krieger, A. M., & Rosenbaum, P. R., (1998). Cornfield's inequality. In P. Armitage and T. Colod (Eds.), *Encyclopedia of Biostatistics* (pp. 959-955). New York, NY: Wiley. doi: 10.1002/0470011815.b2a03040

Golinelli, D., Ridgeway, G., Rhoades, H., Tucker, J., & Wenzel, S. (2012). Bias and variance trade-offs when combining propensity score weighting and regression: With an application to HIV status and homeless men. *Health Services Outcomes Research Methodology, 12*(2-3), 104-118. doi:10.1007/s10742-012-0090-1

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J. et al. (2011). Effects of adjusting for instrumental variables on bias and precision of

effect estimates. *American Journal of Epidemiology, 174*(11), 1213-1222. doi: 10.1093/aje/kwr364

Pearl, J. (2011). Invited commentary: Understanding bias amplification. *American Journal of Epidemiology, 174*(11), 1223-1227. doi: 10.1093/aje/kwr352

Reshetnyak, E., Cham, H., & Hughes, J. N. (this issue). Invited commentary: Using marginal structural modeling for grade retention effects. *Multivariate Behavioral Research*.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41-55. doi:10.2307/2335942

Steiner, P. M., Park, S., & Kim, Y. (this issue). Invited commentary: Identifying causal estimands for time-varying treatments measured with time-varying (age or grade-based) instruments. *Multivariate Behavioral Research*.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1-21. doi:10.1214/09-STS313

van der Laan, M. J., and Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *International Journal of Biostatistics, 6*(1), 1557-4679. doi: 10.2202/1557-4679.1181

Vandecandelaere, M., Schmitt, E., Vanlaar, G., De Fraine, B., & Van Damme, J. (2015). Effects of kindergarten retention for at-risk children's mathematics development. *Research Papers in Education, 30*(3), 305-326. doi: 10.1080/02671522.2014.919523

Vandecandelaere, M., Vansteelandt S., De Fraine, B., & Van Damme, J. (this issue). Time-varying treatments in observational studies: Marginal structural models of the effects of early grade retention on math achievement. *Multivariate Behavioral Research.*

VanderWeele, T. J., & Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology, 22*(1), 42-52. doi:10.1097/EDE.0b013e3181f74493

Vansteelandt, S., Bekaert, M. and Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*, 21(1), 7-30. doi: 10.1177/0962280210387717